On Bicycles Stolen on Different Premises in Toronto from 2017 to 2023*

Emily Su

16 January 2024

Toronto's population is increasing alongside the number of bicycle thefts across the city. We used data from Open Data Toronto to create a graph to discover any trends and differences between the number of bicycle thefts before 2020 and 2020 onwards. Our analysis found that there was a steep decline in the number of bicycle thefts after 2020 with a majority of thefts occurring in the premise type, Outside, in most years except for 2019, 2020, and 2021. Factors like the pandemic could have influenced the decline however, further investigation is needed on the demographic of people reporting.

Table of contents

*Data and code are available at: https://github.com/moonsdust/toronto-bicycles-stolen

1 Introduction

As Toronto's population increases, roads increasingly become packed with roads and bicycles have become one of the more favourable options for transportation, especially during the summer. However, this means that bicycle theft happens around Toronto and on different premises. With the pandemic starting in 2020, we are interested in seeing how the number of bicycle thefts has changed before 2020 versus 2020 onwards.

In this paper, we will look at the number of bicycles stolen from 2017 to 2023 on different types of premises (Commercial (ex: Gas stations, construction sites, etc.), Outside (ex: Parking lots, roads, etc.), House, Apartment, Transit (ex: subway station), Educational, and Other) in Toronto. We then discuss our results and any implications of it.

2 Process & Data

In this section, we will discuss the process and some of the code used to analyze and clean the data from `Open Data Toronto` on Bicycle Thefts and create a graph from it (Open Data Toronto 2023). See the footnote on the first page for the GitHub repository, which contains the data and all code pertaining to this paper.

2.1 Plan

The sketch showing how we expect the dataset and final graph to look can be found under `inputs/sketches/sketch.png` in the GitHub repository in the footnote of the first page.

2.2 Simulate

We will first add the preamble documentation and setup our workspace. We will use `R (programming language)` (R Core Team 2023), `tidyverse` (Wickham et al. 2019), `janitor` (Firke 2023), `ggplot2` (Wickham 2016), and `knitr` (Xie 2014).

```
#### Preamble ####
# Purpose: To create a graph of the number of bicycles stolen
# in Toronto on different premises from 2017 to 2023 by reading
# in data from Open Data Toronto.
# Author: Emily Su
# Email: em.su@mail.utoronto.ca
# Date: 16 January 2024
# Prerequisites: Know where the data is for bicycle thefts in Toronto.
```

Now, we will create simulated data for the number of bicycle thefts on different Toronto premises from 2017 to 2023. These premises are the following: Commercial, Outside, House, Apartment, Transit, Educational, and Other.

Table 1: Simulated Data of the Number of Bicycle Stolen on Different Premises in Toronto (2017-2023)

premise	year	num_bikes_stolen
Commercial	2017	2
Commercial	2018	4
Commercial	2019	38
Commercial	2020	5
Commercial	2021	44
Commercial	2022	28

2.3 Acquire

In this step, we will first read in data from `Open Data Toronto` on Bicycle Thefts and then clean the dataset (Open Data Toronto 2023).

Next, we will clean the column names of the raw data's dataframe, filter for the year 2017 to 2023, and select the following columns: `occ_year` and `premises_type`. We will then rename the column `occ_year` to `year` and create a new column called `num_bikes_stolen` to contain the number of bikes stolen for each `premises_type` for each `year`.

Now, we will run tests on the clean data to see if it passes the following conditions:

1. "year" does not contain years before 2017 and after 2023
2. We have 7 types of "premises_type"
3. num_bikes_stolen is greater than or equal to 0.

```
# Tests
# 1. "year" does not contain years before 2017 and after 2023
cleaned_bike_data$year |> min() == 2017
```

```
[1] TRUE
```

```
cleaned_bike_data$year |> max() == 2023
```

```
[1] TRUE
```

```
# 2. We have 7 types of "premise"
cleaned_bike_data$premises_type |>
  unique() |>
  length() == 7
```

```
[1] TRUE
```

```
# 3. num_bikes_stolen is greater than or equal to 0.
cleaned_bike_data$num_bikes_stolen |> min() >= 0
```

```
[1] TRUE
```

2.4 Explore

We will now create a table and graph for the cleaned data.

Table 2: Number of Bicycles Stolen in Toronto on Different Premises from 2017 to 2023

year	premises_type	num_bikes_stolen
2017	Outside	1262
2017	Commercial	544
2017	Educational	230
2017	Transit	80
2017	Apartment	815
2017	House	561
2017	Other	425
2018	House	551
2018	Apartment	984
2018	Other	428
2018	Transit	114
2018	Outside	1281
2018	Commercial	465
2018	Educational	169
2019	House	498
2019	Outside	1063
2019	Apartment	1135
2019	Commercial	385
2019	Transit	112
2019	Other	420
2019	Educational	137

year	premises_type	num_bikes_stolen
2020	Apartment	1322
2020	House	670
2020	Other	495
2020	Outside	959
2020	Commercial	342
2020	Educational	86
2020	Transit	45
2021	Educational	74
2021	Commercial	330
2021	Apartment	1019
2021	Outside	796
2021	House	447
2021	Other	454
2021	Transit	63
2022	House	255
2022	Outside	941
2022	Apartment	836
2022	Other	343
2022	Commercial	389
2022	Educational	147
2022	Transit	72
2023	Apartment	700
2023	Outside	775
2023	Commercial	425
2023	Other	477
2023	House	378
2023	Educational	111
2023	Transit	84

3 Results & Discussion

We obtained data from Open Data Toronto on Bicycle Thefts that started in 2013 (Open Data Toronto 2023). We then used the statistical programming language `R` (R Core Team 2023), `tidyverse` (Wickham et al. 2019), and `janitor` (Firke 2023) to clean, tidy, and analyze the dataset. We followed the process (Plan, Simulate, Acquire, and Explore) to obtain a graph on the number of bicycles stolen during each year in Toronto for each type of premise (Figure 1) using `ggplot2` (Wickham 2016), and `knitr` (Xie 2014).

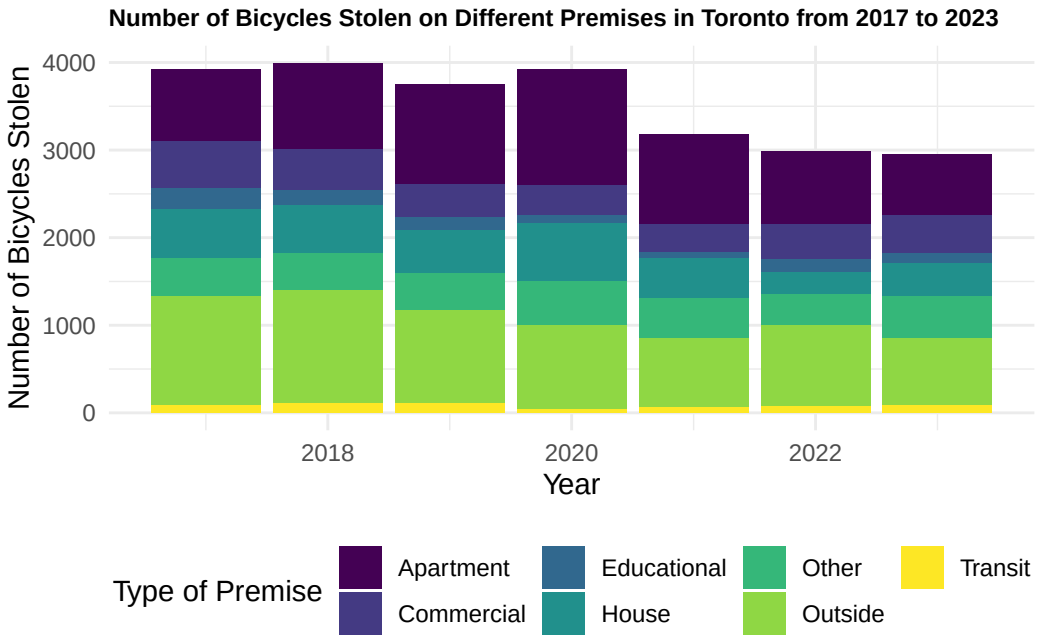

Figure 1: Number of Bicycles Stolen on Different Toronto Premises from 2017 to 2023

Figure 1 from our results shows that 2018 had the most number of bicycle thefts with the most occurring Outside such as a parking lot and road. In contrast, the least number of bicycle thefts occurred in 2023 with the most thefts occurring at the premise type, outside. The number of bicycle thefts declined after 2020 when compared to the years before 2020. We can also see that in 2020 and 2021, the most number of bicycle thefts occurred at an apartment. We can predict that this could have been due to factors like the lockdowns and people staying at home more. This can also been seen in 2019 as well. For all years except for 2019, 2020, and 2021, the most bicycle thefts occurred outside.

However, this data only contained bicycle thefts that were reported and the result we would have gotten if all bicycle thefts were reported could look completely different from our current results. Some might have not reported due to various reasons such as historical discrimination from the police, the police station being located too far, or not having access to a phone. Further investigation is needed on the demographics (such as income) of people reporting bicycle theft.

References

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Open Data Toronto. 2023. *Bicycle Thefts*. Toronto: Open Data Toronto. https://open.toronto.ca/dataset/bicycle-thefts/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.